Martin Popel, Roman Sudarikov, Ondřej Bojar, Rudolf Rosa, and **Jan Hajič**

{popel,sudarikov,bojar,rosa,hajic}@ufal.mff.cuni.cz

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

TectoMT

a deep-linguistic core of the combined Chimera MT system

Building on multilingual standards

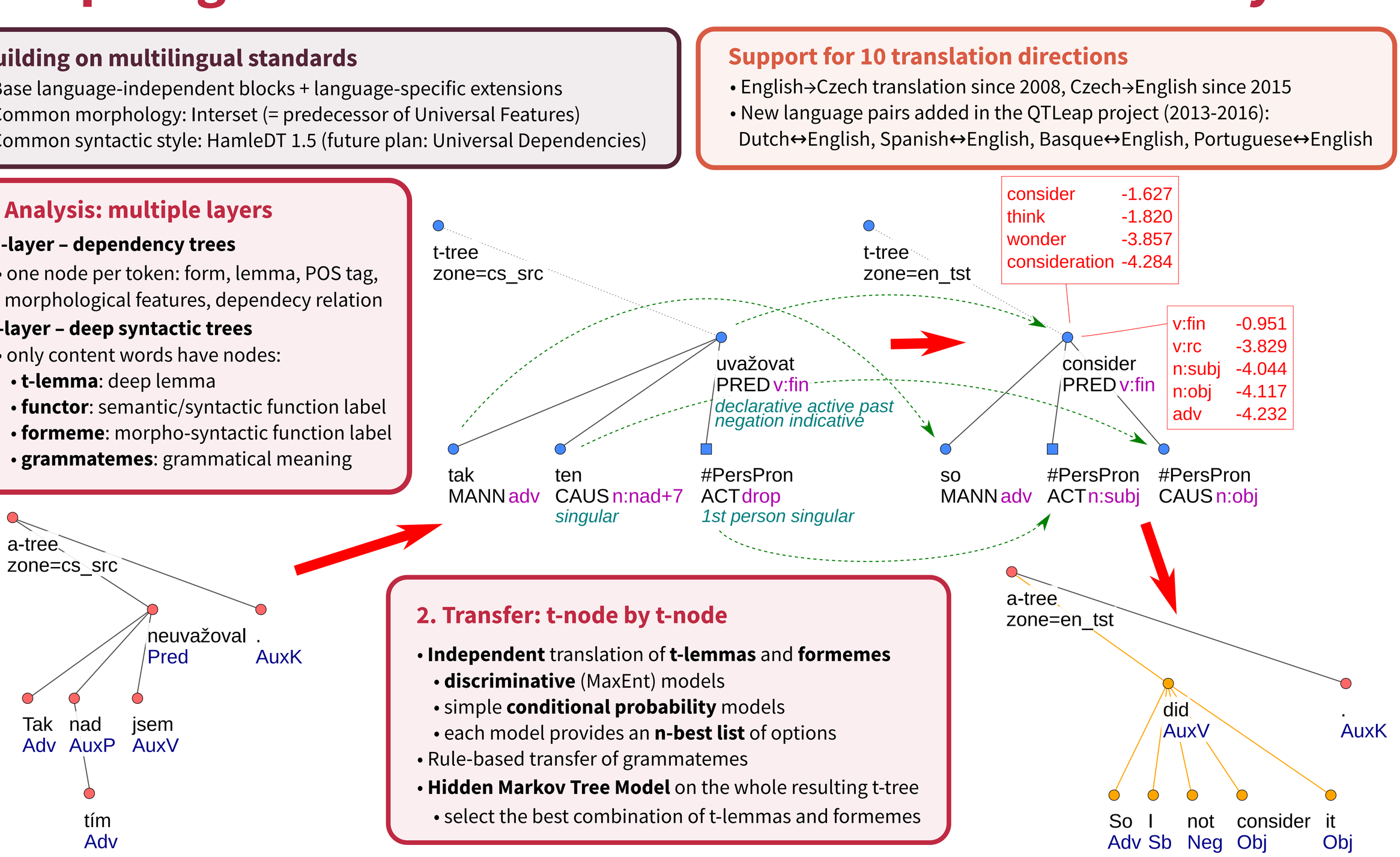
- Base language-independent blocks + language-specific extensions
- Common morphology: Interset (= predecessor of Universal Features)
- Common syntactic style: HamleDT 1.5 (future plan: Universal Dependencies)

Support for 10 translation directions

- English→Czech translation since 2008, Czech→English since 2015
- New language pairs added in the QTLeap project (2013-2016):
  Dutch↔English, Spanish↔English, Basque↔English, Portuguese↔English

1. Analysis: multiple layers

- **a-layer – dependency trees**
  - one node per token: form, lemma, POS tag, morphological features, dependecy relation
- **t-layer – deep syntactic trees**
  - only content words have nodes:
    - **t-lemma**: deep lemma
    - **functor**: semantic/syntactic function label
    - **formeme**: morpho-syntactic function label
    - **grammatemes**: grammatical meaning

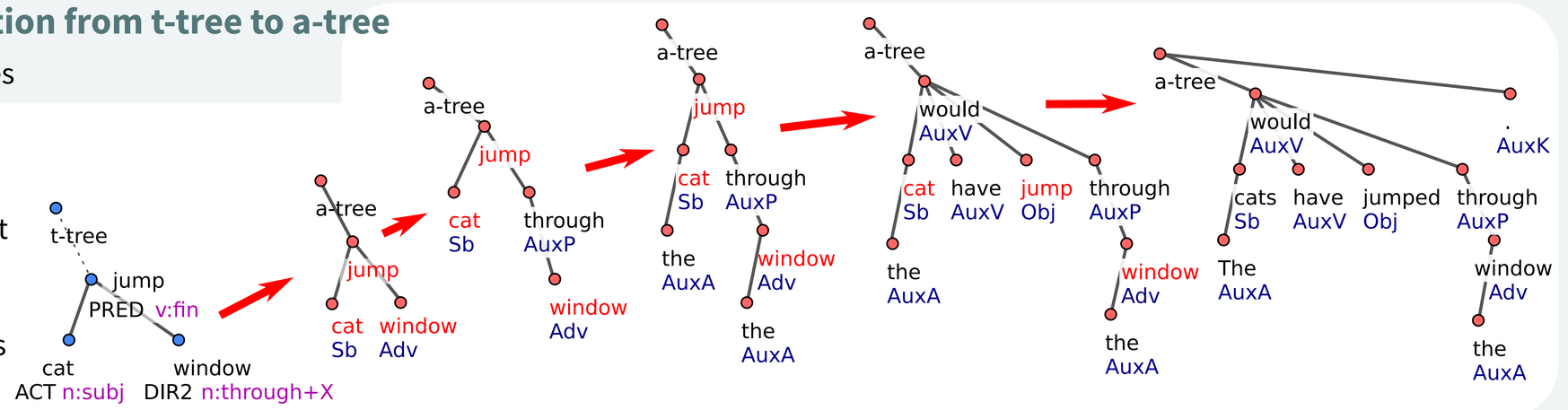


2. Transfer: t-node by t-node

- **Independent** translation of **t-lemmas** and **formemes**
  - **discriminative** (MaxEnt) models
  - simple **conditional probability** models
  - each model provides an **n-best list** of options
- Rule-based transfer of grammatemes
- **Hidden Markov Tree Model** on the whole resulting t-tree
  - select the best combination of t-lemmas and formemes

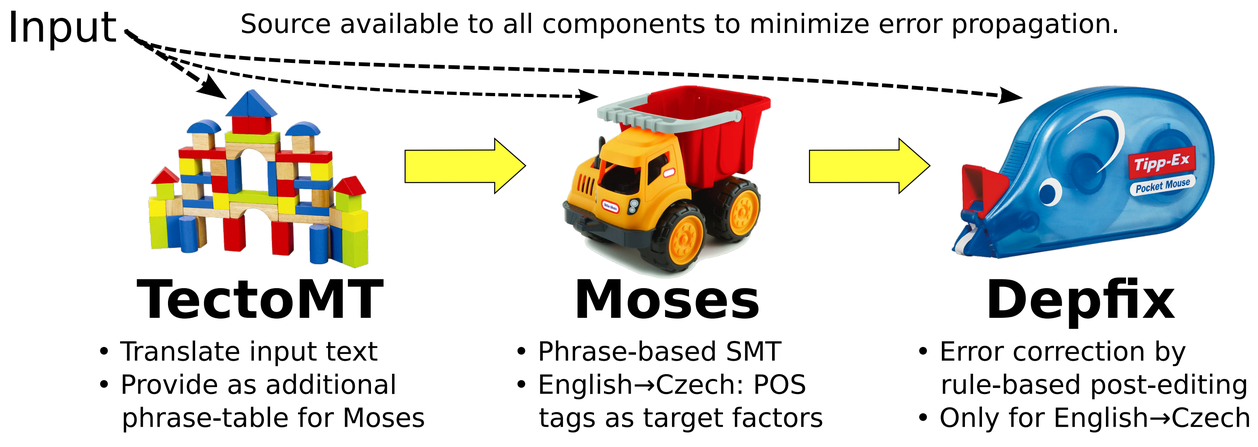
3. Synthesis: gradual transformation from t-tree to a-tree

- Grammatemes → morphological attributes
- Enforce basic word order
- Enforce subject-predicate agreement
- Formemes → prepositions, conjunctions
- Add aux verbs, remove imperative subject
- Add articles, negation particles
- Inflection (MorphoDiTa + Flect)
- Delete repeated prepositions in conjuncts
- Punctuation, capitalization



Chimera

TectoMT + Moses (+ Depfix)



TectoMT
- Translate input text
- Provide as additional phrase-table for Moses

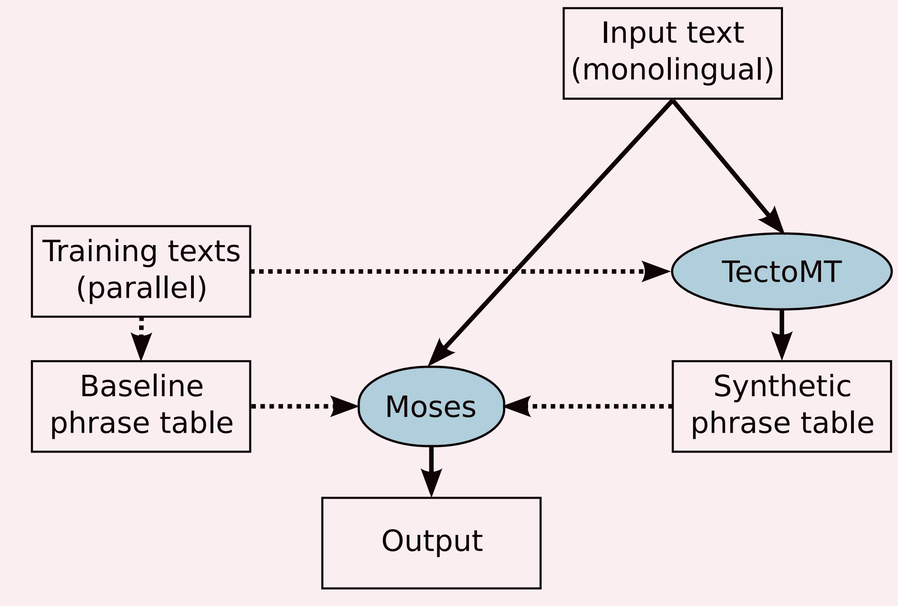
Moses
- Phrase-based SMT
- English→Czech: POS tags as target factors

Depfix
- Error correction by rule-based post-editing
- Only for English→Czech

Beneficial combination of conceptually different systems

- Moses by itself better than TectoMT - but **even better when combined with TectoMT**
- TectoMT provides many novel word forms (unseen in parallel training data)
- TectoMT translations have better grammatical coherence
  - kept in long phrases in the phrase table, since it matches the input

Two-headed Chimera



Find TectoMT and Chimera online!

- http://ufal.cz/tectomt
- http://ufal.cz/chimera

The development is currently supported by the QTLeap 7th FP project No. 610516 (http://qtleap.eu). This work has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).